DOCUMENT RESUME

ED 480 905 TM 035 258

AUTHOR Mislevy, Robert J.

TITLE Argument Substance and Argument Structure in Educational Assessment. CSE Report.

INSTITUTION National Center for Research on Evaluation, Standards, and Student Testing, Los Angeles, CA.; California Univ., Los Angeles. Center for the Study of Evaluation.

SPONS AGENCY Institute of Education Sciences (ED), Washington, DC.; National Science Foundation, Arlington, VA.

REPORT NO CSE-R-605

PUB DATE 2003-09-00

NOTE 38p.; Paper presented at the Conference on Inference, Culture, and Ordinary Thinking in Dispute Resolution (New York, NY April 27-29, 2003).

CONTRACT R305B960002;REC-0129331

PUB TYPE Reports - Descriptive (141) -- Speeches/Meeting Papers (150)

EDRS PRICE EDRS Price MF01/PC02 Plus Postage.

DESCRIPTORS *Educational Assessment; Educational Philosophy; Elementary Secondary Education; *Student Evaluation

IDENTIFIERS *Paradigm Argument

ABSTRACT

Educational assessment is reasoning from observations of what students do or make in a handful of particular circumstances, to what they know or can do more broadly. Practice has changed a great deal over the past century, in response to evolving conceptions of knowledge and its acquisition, views of schooling and its purposes, and technologies for gathering and evaluating response data. Conceptions of what constitutes assessment data, how it should be interpreted, and what kind of inferences are to be drawn differ radically when cast under different psychological perspectives. If the structure of assessment arguments is distinguished from their substance, greater continuity is seen. Developments in this area have been more in the nature of extension, elaboration, refinement, and explication of argument structures, as they have been prompted by more radical changes in culture and substance. (Contains 13 figures and 51 references.) (Author/SLD)

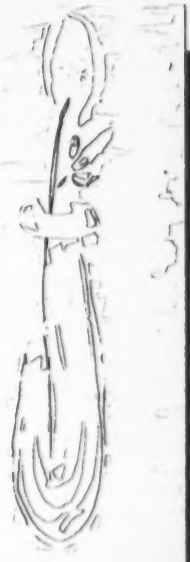
Reproductions supplied by EDRS are the best that can be made from the original document.

CRESST
National Center for Research on Evaluation, Standards, and Student Testing

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
K. Hurst
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
1

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

ED 480 905



Argument Substance and Argument Structure in Educational Assessment

CSE Report 605

Robert J. Mislevy
CRESST/University of Maryland

TM035258

UCLA Center for the Study of Evaluation
In Collaboration With:
University of California, Los Angeles • Graduate School of Education • The RAND Corporation
University of Southern California • Educational Testing Service
University of Pittsburgh • University of Colorado

BEST COPY AVAILABLE

Argument Substance and Argument Structure in Educational Assessment

CSE Report 605

Robert J. Mislevy
CRESST/University of Maryland

September 2003

Center for the Study of Evaluation
National Center for Research on Evaluation,
Standards, and Student Testing
Graduate School of Education & Information Studies
University of California, Los Angeles
Los Angeles, CA 90095-1522
(310) 206-1532

Project 3.6 Group Activity on Cognitive Validity
Robert J. Mislevy, Project Director, CRESST/University of Maryland

Copyright © 2003 The Regents of the University of California

Support was provided by the Educational Research and Development Centers Program, PR/Award Number R305B960002, as administered by the Institute of Education Sciences, U.S. Department of Education. The PADI project is supported by the National Science Foundation under grant, REC-0129331 (PADI Implementation Grant).

The findings and opinions expressed in this report do not reflect the positions or policies of the National Institute on Student Achievement, Curriculum, and Assessment, the Institute of Education Sciences, or the U.S. Department of Education.

ARGUMENT SUBSTANCE AND ARGUMENT STRUCTURE IN EDUCATIONAL ASSESSMENT[1]

Robert J. Mislevy

CRESST/University of Maryland

Abstract

Educational assessment is reasoning from observations of what students do or make in a handful of particular circumstances, to what they know or can do more broadly. Practice has changed a great deal over the past century, in response to evolving conceptions of knowledge and its acquisition, views of schooling and its purposes, and technologies for gathering and evaluating response data. Conceptions of what constitutes assessment data, how it should be interpreted, and what kind of inferences are to be drawn differ radically when cast under different psychological perspectives. If we distinguish the structure of assessment arguments from their substance, we see greater continuity. Developments here have been more in the nature of extension, elaboration, refinement, and explication of argument structures, as they have been prompted by more radical changes in culture and substance.

Introduction

Educational assessment is reasoning from observations of what students do or make in a handful of particular circumstances, to what they know or can do more broadly. Practice has changed a great deal over the past century, in response to evolving conceptions of knowledge and its acquisition, views of schooling and its purposes, and technologies for gathering and evaluating response data. It is not merely that forms of data have changed over the years. Conceptions of what constitutes assessment data, how it should be interpreted, and what kind of inferences are to be drawn differ radically when cast under different psychological perspectives, including prominently those known as trait or

---

[1] Presented at Conference on Inference, Culture, and Ordinary Thinking in Dispute Resolution, Benjamin N. Cardozo School of Law, Yeshiva University, New York, New York, April 27-29, 2003. This work builds on research with Linda Steinberg and Russell Almond at Educational Testing Service on the structure of educational assessments. We gratefully acknowledge the influence of David Schum's investigations into evidentiary reasoning on our thinking. Thanks to Geneva Haertel for comments on an earlier version.

differential, behavioral, information-processing, and sociocultural (Greeno, Collins, & Resnick, 1997; National Research Council, 2001).

Not everything is different, though. Educational assessment is a special case of evidentiary reasoning, which in turn is a special case of argument. If we distinguish the structure of assessment arguments from their substance, we see greater continuity in the forms of data, their interpretation, and the inferences drawn from them. Developments have been more in the nature of the extension, elaboration, refinement, and explication of assessment argument structures, as they have been prompted by each succeeding wave of ambitions in the design and use of assessments. We see accumulation and elaboration of recurring themes and relationships—problems of reasoning from limited numbers of observations, for example, and basing inference on the reports of imperfect raters.

The structure of educational assessments can be understood in terms of concepts and representational forms for arguments introduced by Wigmore (1937) and Toulmin (1958), broadened and extended more recently by contemporary evidence scholars such as Schum (1994), Tillers (Tillers & Schum, 1991), and Anderson and Twining (1991). These ideas fit well with the contemporary conception of test validity as the grounding of the argument and the quality of the evidence for inferences or decisions based on students' performances (Cronbach & Meehl, 1955; Embretson, 1983; Kane, 1992; Messick, 1989).

This presentation begins with a brief review of Toulmin's argument structure, including the role of claims, data, warrants, and qualifiers. This structure is related to assessment arguments, as understanding of them has evolved in the educational and psychological measurement literature. We then consider four psychological perspectives from which assessment arguments might be cast. We see how the psychological perspectives impact the nature of claims, evidence, warrants, and qualifiers in assessment, and recognize the elaborations of the basic structure that are needed to accommodate increasingly sophisticated arguments.[2]

---

[2] The reader interested in the role of arguments in assessments as they relate to assessment design and delivery systems more generally is referred to Almond, Steinberg, and Mislevy (2002) and Mislevy, Steinberg, & Almond (2003); and Mislevy, Wilson, Ercikan, and Chudowsky (2003) for the connection to psychometric modeling.

Toulmin's Argument Structure

Philosopher Stephen Toulmin (1958) provided terminology for talking about how we use substantive theories and accumulated experience (say, about algebra and how kids learn it) to reason from particular data (Joe's solutions) to a particular claim (what Joe understands about algebra). Figure 1 outlines the structure of a simple argument. The *claim* is a proposition we wish to support with *data*. The arrow represents inference, which is justified by a *warrant*, a generalization that justifies the inference from the particular data to the particular claim. Theory and experience—both personal and formal, such as empirical studies and prior research findings—provide *backing* for the warrant. In any particular case we reason back through the warrant, so we may need to qualify our conclusions because there may be *alternative explanations* for the data.

In practice, of course, an argument and its constituent claims, data, warrants, backing, and alternative explanations will be more complex than Figure 1. An argument often consists of many propositions and data elements, involves chains of reasoning, and often contains dependencies among claims and various pieces of data. Wigmore's (1937) earlier system of charting (modernized by Anderson & Twining, 1991), accommodates elaborations, and includes ideas such as chaining and conjunction that turn out to be useful in assessment. A further extension that is central to educational assessment is the use of statistical models as one aspect of a warrant (Schum 1994, Section 4.5). Formal assessment applications employ stochastic models for specified qualities in students' observable performances (e.g., correct answers, coherent essays, or space-splitting moves in troubleshooting), as a function of variables that characterize knowledge and skill—however conceived (Mislevy, 1994).

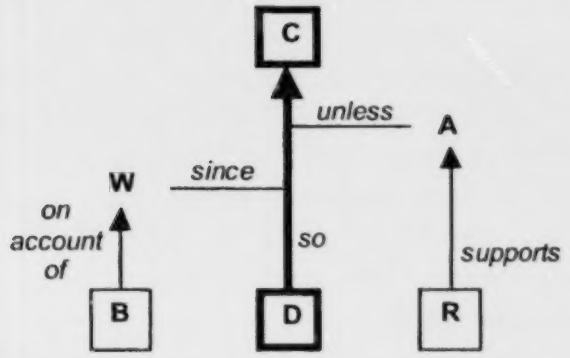


Figure 1. Toulmin's (1958) structure for arguments. Reasoning flows from *data* (D) to *claim* (C) by justification of a *warrant* (W), which in turn is supported by *backing* (B). The inference may need to be qualified by *alternative explanations* (A), which may have *rebuttal evidence* (R) to support them.

The history of test theory in the 20th century is a steady march toward an explication of its foundations in evidentiary reasoning, starting from a collection of practically useful testing techniques that commingled notions of psychology, method, and purpose. By 1961, Harold Gulliksen, speaking at the 25th anniversary of the Psychometric Society, was able to describe "the central problem of test theory" as "the relation between the ability of the individual and his [or her] observed score on the test" (Gulliksen, 1961). Twenty-five years later, at the 50th anniversary meeting, Charles Lewis observed that "much of the recent progress in test theory has been made by treating the study of the relationship between responses to a set of test items and a hypothesized trait (or traits) of an individual as a problem of statistical inference" (Lewis, 1986). In his influential chapter on *validity* in the fifth edition of *Educational Measurement*, Messick (1989) described this most central concept in measurement as "an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the *adequacy* and *appropriateness* of *inferences* and *actions* based on test scores or other modes of assessment" (p. 13).

What are the essential elements of an assessment argument? Another quotation from Messick provides a good starting point:

> A construct-centered approach [to assessment design] would begin by asking what complex of knowledge, skills, or other attribute should be assessed, presumably because they are tied to explicit or implicit objectives of instruction or are otherwise valued by society. Next, what behaviors or performances should reveal those constructs, and what

> tasks or situations should elicit those behaviors? Thus, the nature of the construct guides the selection or construction of relevant tasks as well as the rational development of construct-based scoring criteria and rubrics.
>
> Messick, 1994, p. 16.

Note the focus on *structure* rather than *substance*. We will be able to identify these central elements of assessment design with elements of Toulmin's argument structures. Further, the essential sources of invalidity Messick identified, "construct-irrelevant variance" and "construct underrepresentation," correspond to kinds of alternative explanations for poor and good performance, other than the targeted knowledge or skill.

Four Psychological Perspectives

An oft-stated axiom in evidentiary reasoning is that data are not evidence until their relationship to some conjecture, some claim, is established (Schum, 1987, p. 16). In any domain of reasoning, knowledge, beliefs, experience, and practices are the source of claims, data, warrants, and alternative explanations. In educational assessment, it is one's belief about the nature and acquisition of knowledge that shapes the why and the what of evidentiary reasoning. This section outlines four perspectives on knowledge and learning under which instruction and assessment might be cast (GCR). These perspectives differ in terms of levels of description and focus of attention, with respect to patterns of acquiring and using knowledge. Naturally this taxonomy is overly simple; there are substantial variation in beliefs and approaches among researchers from any of these perspectives, and practical assessment generally requires viewing students' learning from multiple perspectives jointly.[3] Nevertheless, drawing sharp distinctions among perspectives will allow us to see clearly the implications that different psychological stances hold for assessment arguments.

- A *trait* perspective. Messick (1989, p. 15) defines a trait as "a relatively stable characteristic of a person—an attribute, enduring process, or disposition—which is consistently manifested to some degree when relevant, despite considerable variation in the range of settings and

[3] People who are doing research have the luxury of being able to pick which of the myriad aspects of learning they want to focus on. People who are learning, and people who are helping them learn, don't.

circumstances." Hypothetical (hence, inherently unobservable) numbers are proposed to locate people along continua of mental characteristics, just as their heights and weights locate them along continua of physical characteristics. The interest in people's differential status on common traits, useful in selection, prediction, and educational decisions, explains why this perspective is also called "differential" psychology.

- A *behaviorist* perspective. The focus is on targeted behavior in a domain of relevant situations, as both the behavior and the situation are viewed by the assessor. Knowledge is the organized accumulation of stimulus-response associations, which serve as the components of skills. People learn by acquiring simple components of a skill, then acquiring more complicated units that combine or differentiate the simpler units. Stimulus-response associations can be strengthened by reinforcement or weakened by inattention. Domains of knowledge can be analyzed in terms of the component information, skills, and procedures to be acquired.

- An *information-processing* perspective. Epitomized in Newell and Simon's (1972) landmark volume *Human Problem Solving*, the information-processing perspective examines the procedures by which people acquire, store, and use knowledge to solve problems. Strong parallels to computation and artificial intelligence appear in the use of rules, production systems, task decompositions, and means-ends analyses. The key insight is modeling problem solving in these terms in light of the capabilities and the limitations of human thought and memory that are revealed by psychological experiments.

- A *sociocultural* perspective. A situative or sociocultural perspective stresses how the knowledge is conditioned and constrained by the technologies, information resources, representation systems, and social situations with which they interact. The situative perspective incorporates explanatory concepts that have proved useful in fields such as ethnography and sociocultural psychology to study "collaborative work, ... mutual understanding in conversation, and other characteristics of interaction that are relevant to the functional success of the participants' activities" (Greeno, et al., 1997, p. 7).

Psychological Perspectives and Assessment Arguments

What does a psychological perspective provide for an assessment argument? Everything, basically; it determines the nature of every element in Toulmin's argument structure and the rationale that orchestrate them as a coherent argument. A psychological perspective provides a universe of discourse

for assessment: What kinds of things one might say concerning students (claims), what kinds of things one wants to see (data), and why the two are related in the first place (warrants). There are always at least two classes of data in an assessment argument: aspects of the circumstances in which the student is acting, over which an assessment designer generally has principal influence, and aspects of the student's behavior in the situations, over which the student has principal influence. Additional knowledge about the student's history or relationship to the observational situation may be further required. These latter factors are essential in assessment in practice, even though they are often tacit, embedded in familiar forms and practices. The traditions of the psychological perspective also determine what counts as backing for warrants, and the kinds of alternative explanations for performance that constitute threats to the argument.

Assessment Arguments Under the Trait / Differential Perspective

Many familiar tools of assessment began to evolve at the dawn of the 20th century under the perspective of trait psychology, initially in a quest to "measure people's intelligence." Under trait psychology, claims about students are phrased in terms of their status on unobservable traits. What constitutes observable evidence about traits? When Charles Spearman used scores on a fixed set of knowledge and puzzle-solving tasks to "measure intelligence," the notion of a trait was not new. Paul Broca had attempted to assess "intelligence" in the previous century by charting cranial volumes, as had Francis Galton by measuring reaction times. The idea of observing behavior in samples of standardized situations wasn't new either. Three thousand years earlier, the Chinese discovered that observing an individual's performance under controlled conditions could support predictions of performance under broader conditions over a longer period of time (Wainer, et al., 2000, p. 2). The essence of mental measurement under trait psychology was a confluence of these concepts: identifying "traits" with tendencies to behave in prescribed ways in these prescribed situations.

The conjoining of this psychological perspective and methodological tools suited the mass educational system that also arose in the United States at the turn of the century (Glaser, 1981). Educators were attempting to select or place large numbers of students into instructional programs, but had limited resources to gather information about each student, offer many options, or tailor programs

to students once a decision was made. This decision-making context encouraged building assessment systems around a small number of broadly construed and widely applicable student characteristics, stable over time and informed by data that were easy to gather and summarize.

From the presumption that a given trait influences behavior over a wide variety of situations, it follows that observations over a wide range of situations can provide evidence about that trait. In fact, writing in the context of measuring intelligence ("g" in his notion), Spearman posited his "theorem of indifference of the indicator":

> This means that, for the purpose of indicating the amount of g possessed by a person, any test will do just as well as any other, provided only that its correlation with g is equally high. With this proviso, the most ridiculous "stunts" will measure the self-same g as will the highest exploits of logic or flights of imagination.
>
> Another consequence of the indifference of the indicator consists in the significance that should be attached to personal estimates of "intelligence" made by teachers and others. However unlike may be the kinds of observation from which these estimates may have been derived, still insofar as they have a sufficiently broad basis to make the influence of g dominate over that of the s's [subjects], they will tend to measure precisely the same thing.
>
> And here, it should be noticed, we come at last upon the secret why all the current tests of "general intelligence" show high correlation with one another, as also with g itself. The reason lies, not in the theories inspiring these tests (which have been most confused), nor in the uniformity of construction (for this has often been wildly heterogeneous), but wholly and solely in the above-shown "indifference of the indicator." Indeed, were it worth while, tests could be constructed which had the most grotesque appearance, yet after all would correlate quite well with all the others.
>
> Spearman, 1927, pp. 197-198.

Pet Shop Display (Figure 2) is an example of an "analytical reasoning" task, an item type used in the Law School Admissions Test (LSAT) and, until October 2002, in the GRE. The description of analytic reasoning items from the LSAT's

web site[4] clearly takes a trait perspective: "Analytical reasoning items are designed to measure the ability to understand a structure of relationships and to draw conclusions about the structure." Such items are included in the LSAT not because either lawyers or law students routinely have to solve problems just like these in their jobs or their studies, but because there is empirical evidence that students who can solve these kinds of puzzles tend to perform better in law school than students who can't. In Toulmin's terms, this is backing for a warrant. The warrant is cast in terms of a trait labeled analytical reasoning: The higher a student's level of analytic reasoning, the more likely the student is to provide correct answers to tasks like these.

Pet Shop Display

Arturo is planning the parakeet display for his pet shop. He has five parakeets, Alice, Bob, Carla, Diwakar, and Etria. Each is a different color; not necessarily in the same order, they are white, speckled, green, blue, and yellow. Arturo has two cages. The top cage holds three birds, and the bottom cage holds two. The display must meet the following additional conditions:

Alice is in the bottom cage.
Bob is in the top cage and is not speckled.
Carla cannot be in the same cage as the blue parakeet.
Etria is green.
The green parakeet and the speckled parakeet are in the same cage.

1. If Carla is in the top cage, which of the following must be true?

a) The green parakeet is in the bottom cage.
b) The speckled parakeet is in the bottom cage.
c) Diwakar is in the top cage.
d) Diwakar is in the bottom cage.
e) The blue parakeet is in the top cage.

Figure 2. An analytical reasoning item. A typical analytical reasoning item begins with a description of a situation with interrelated entities, properties, and relationships. One or more questions are posed that ask about further properties of the situation that are implied by the initial conditions.

Figure 3 is the structure of the argument that leads from observing Sue give a correct answer to the Pet Shop Display to the claim that she has a high level of analytical reasoning ability. (We will address below more refined claims in terms of the values of a continuous analytic reasoning variable.) Two data elements are

[4] http://www.lsac.org/qod/questions/analytical.htm (downloaded February 26, 2003)

shown, namely the item content that satisfies the qualities stated generally in the definition of analytic reasoning and her response in that situation. There are issues of control and sequence here. The assessor was responsible for the first element when the item was presented to Sue. She could respond either correctly or incorrectly, and the basic structure of the argument would be complete. Note that the data elements and the claim are specifically about Sue, while the warrant is a broader generalization that presumably justifies inference about Sue as an instance thereof.

The item content and student performance data elements in Figure 3 should be modeled in greater detail to reflect an important feature of assessment: Actually neither the situation nor the performance in and of themselves directly constitute the data for the argument, but rather salient aspects of them, as they are perceived by the assessor. The Messick quotation makes clear that these determinations are made in light of the purpose of the assessment and through a perspective on knowledge. For example, a German chemistry major's English-

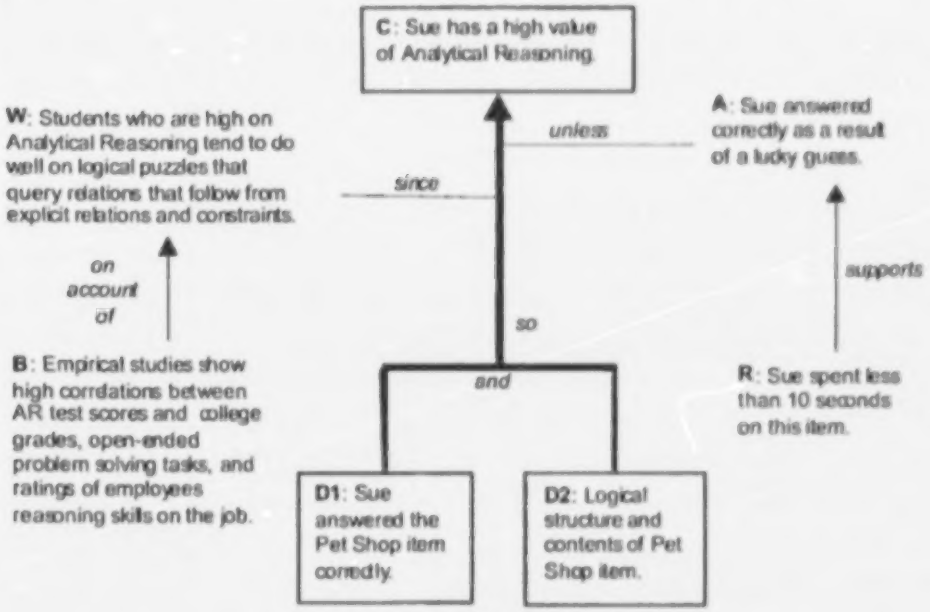

C: Sue has a high value of Analytical Reasoning.
W: Students who are high on Analytical Reasoning tend to do well on logical puzzles that query relations that follow from explicit relations and constraints.
unless
A: Sue answered correctly as a result of a lucky guess.
since
on account of
supports
so
B: Empirical studies show high correlations between AR test scores and college grades, open-ended problem solving tasks, and ratings of employees reasoning skills on the job.
and
R: Sue spent less than 10 seconds on this item.
D1: Sue answered the Pet Shop item correctly.
D2: Logical structure and contents of Pet Shop item.

Figure 3. Toulmin diagram for reasoning from Sue's correct response to her Analytical Reasoning ability. Note that the warrant requires a conjunction of data about the nature of Sue's performance and the nature of the performance situation.

language paragraph on combustion might be evaluated for language control in English class, ignoring the chemistry context and content, but evaluated for scientific accuracy in Chemistry class, ignoring the mechanics of the language. Such considerations constitute warrants for reasoning from unique performances and performance situations to the data for the core assessment argument, as shown in Figure 4.[5] Whenever humans make determinations

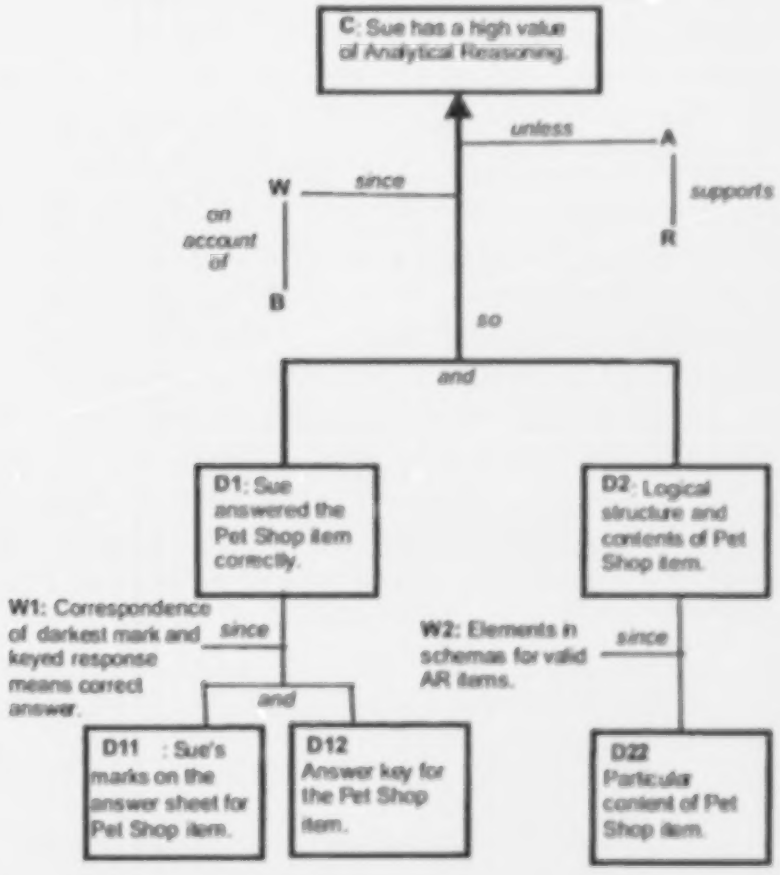

C: Sue has a high value of Analytical Reasoning.
unless
A
W
since
supports
on account of
R
B
so
and
D1: Sue answered the Pet Shop item correctly.
D2: Logical structure and contents of Pet Shop item.
W1: Correspondence of darkest mark and keyed response means correct answer.
since
W2: Elements in schemas for valid AR items.
since
and
D11 : Sue's marks on the answer sheet for Pet Shop item.
D12 Answer key for the Pet Shop item.
D22 Particular content of Pet Shop item.

Figure 4. Elaborated Toulmin diagram for Pet Shop item. Adds detail to the process of reasoning from Sue's performance to the correctness of her answer and from the particulars of the Pet Shop item to its capability to evoke evidence about Analytical Reasoning ability. Note that a proposition such as D1 (Sue answered correctly) can be both a claim that depends on preceding propositions and (provisionally) an element of data for a subsequent claim.

[5] Note that in this figure and those that follow, we may abbreviate or omit from a given Toulmin diagram those elements that are not central to the point that is being made.

BEST COPY AVAILABLE

such as these, questions of sensitivity and objectivity appear just as they do in witness testimony in jurisprudence (Schum, 1994, p. 101 ff.). These considerations introduce alternative explanations at this stage in the full assessment argument for apparent high or low performance. Much effort in educational measurement has gone into both statistical methods and support mechanisms for monitoring and improving the evaluation of performances.

Because no single performance provides conclusive evidence about what a student knows and can do as more generally construed, most educational assessments consist of multiple observations. Figures 5 and 6 shows two ways of depicting an argument with more than one observation. Figure 5 suits a test comprised of several analytical reasoning items, each differing in particulars but all following the same general form and requiring the same kind of reasoning. A single warrant is shown encompassing all of them. Figure 6 suits a situation Spearman described: Inference about Sue's analytical reasoning ability from diverse forms of evidence, including her Pet Shop response, a teacher recommendation, and a grade in algebra class. The justification for each of these is sufficiently distinct to require its own warrant. These elaborations extend beyond the basic Toulmin diagram and move in the direction of Wigmore's diagrams. Wigmore allowed for multiple strands of argumentation, and hierarchies, even webs, of claims and evidence for and against them. Wigmore was particularly interested in recurring patterns of relationships among claims and evidence, which once understood could be recognized and brought to bear on problems that might appear quite different on the surface.

Whenever there are multiple pieces of evidence, often in conflict, sometimes overlapping, the challenge becomes synthesizing their import into a final conclusion. Neither Toulmin nor Wigmore proposed a mechanism to accomplish this. Developments in probability-based reasoning since the 1980s have provided a solution, in the form of Bayesian inference networks (e.g., Jensen, 1996; Edwards, 1998). A Bayes net embeds a substantive argument such as these examples in a joint probability distribution of variables. Claims and data become variables in the network, and qualitative warrants are the starting point for quantitative expressions of relationships between claims and data.

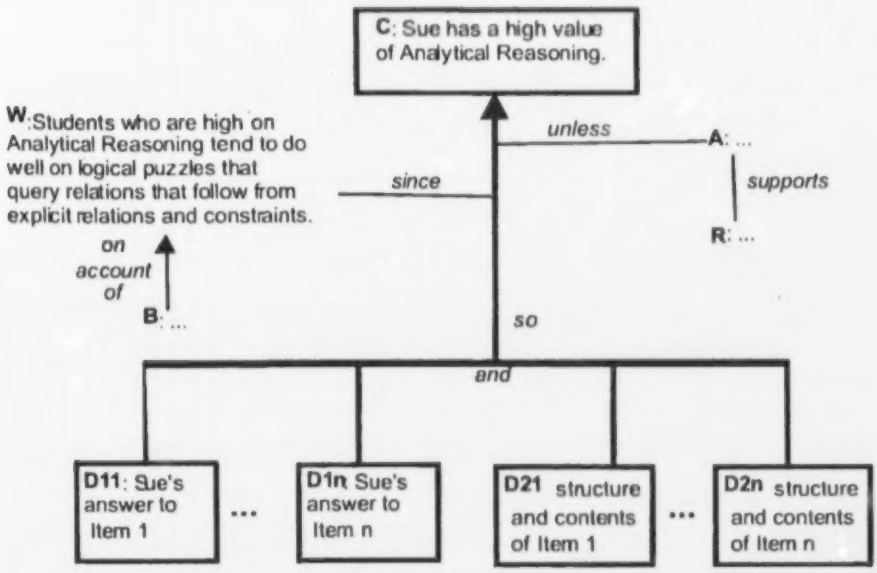

C: Sue has a high value of Analytical Reasoning.
W: Students who are high on Analytical Reasoning tend to do well on logical puzzles that query relations that follow from explicit relations and constraints.
unless
A: ...
supports
R: ...
since
on account of
B: ...
so
and
D11: Sue's answer to Item 1
D1n Sue's answer to Item n
D21 structure and contents of Item 1
D2n structure and contents of Item n

Figure 5. Elaborated Toulmin diagram for multiple pieces of evidence of the same kind about Analytic Reasoning. The same general warrant is employed, as adapted to the particulars of each piece of data as they fit into the same scheme.

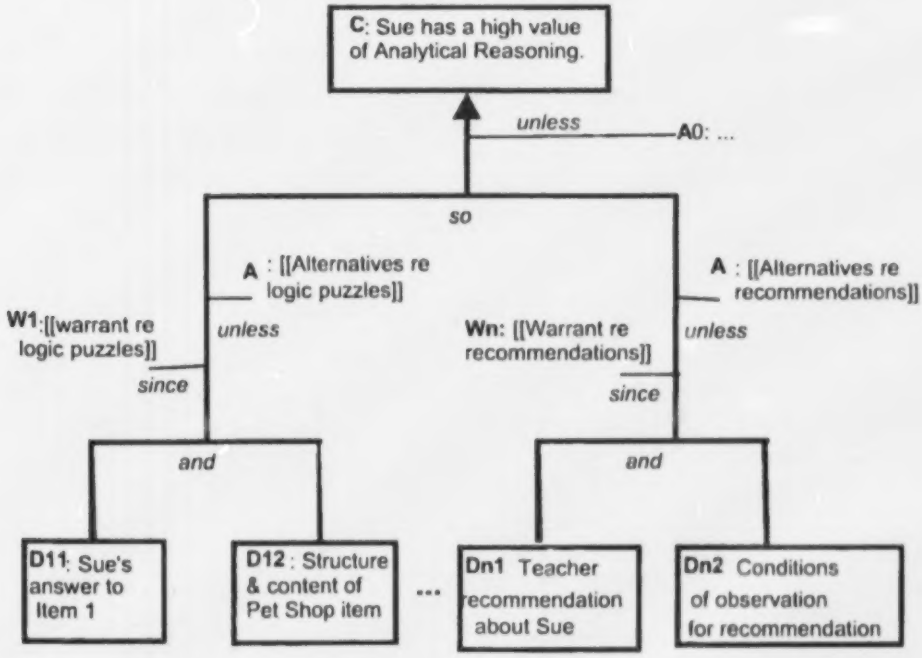


Figure 6. Elaborated Toulmin diagram for multiple pieces of evidence of different kinds about Analytic Reasoning. Different warrants are needed to justify each different kind of data as evidence about Analytical Reasoning.

Psychometric models are special cases of this kind of reasoning. Although the role of probability-based reasoning in assessment is not the focus of this presentation, a few words on the topic are in order (see Mislevy, 1994, and Mislevy & Gitomer, 1996, for more extended discussions). In trait-based applications, traits are unobservable variables that characterize students. Aspects of students' responses are observable variables, which are modeled as depending in probability on the student variables. This is an expression of the warrant in a deductive direction, or expectations for what observables might be if student variables were known to take any particular value. A claim is now expressed as a probability distribution that represents degree-of-belief about the values of the student variables; this belief may in turn be used to support claims about future performance, characterization of knowledge, or likeliness to benefit from such-and-such instructional experience. Observations are generally collected in such a manner as to render observable variables within a given task independent of observations from other tasks, conditional on the (unknown) values of the student variables.

Figure 7 illustrates the probability model for the similar-tasks example.

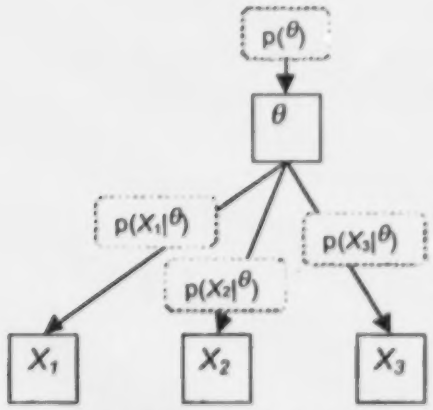


Figure 7. An acyclic direct graph for a statistical model for a test composed of multiple, conditionally independent, Analytical Reasoning items. Student's value of the unobservable Analytic Reasoning variable is denoted by θ; response to Item j is denoted by X_j, 1 if right and 0 if wrong; $p(\theta)$ is a distribution expressing what is known about θ *before* item responses are observed; $p(X_j|\ \theta)$ is a conditional probability distribution for the response to Item j given any particular value of θ. An updated probability distribution $p(\theta\ |\ x_1, \ldots, x_n)$ expressing what is known about θ *after* observing a student's responses is obtained via Bayes theorem.

Once such a model is fit and parameters have been estimated from initial data (pretest data, or 'calibration' data), Bayes theorem can be used to update belief about student variables in light of task performances. The probability model has become an additional aspect of a compound warrant, which permits quantitative expression of belief and the calculus of probability to synthesize multiple, possibly conflicting, possibly overlapping, pieces of evidence. These advantages are not free, of course. First is the additional backing that is required for the probability-based aspects of the warrant, in the form of the pretest data. Further, additional alternative explanations for good or poor performance are introduced in connection with model misspecification and data errors.

As noted above, the contemporary view of test validation concerns examining the support for and potential threats to inferences based on assessment data. Alternative explanations that arise in trait-based assessments address the scope of the trait in question. Does performance in the assessment tasks fail to show the hypothesized relationships with some students' performances due to measurement error? That is, would another try at a different but equivalent set of items point in the other direction? This is an alternative explanation from within the trait perspective. Might some students be solving, say, purported spatial reasoning tasks using non-spatial strategies (French, 1965)? This is an alternative explanation associated with the information-processing perspective. Do the relationships hold for some examinees but not others, as when recent immigrants were deemed unintelligent when their low scores on IQ tests could be explained by lack of familiarity with their new home (Gould, 1981, Chap. 5)? This is an alternative explanation associated with the sociocultural perspective.

As useful as trait-based assessment scores might be for the purposes of selection, classification, certification, or program evaluation, it was ultimately their limitations for the purpose of guiding instruction that led to the rise of assessment from alternative psychological perspectives.

Assessment Arguments Under the Behavioral Perspective

As useful as trait-based tests may be for making placement and selection decisions in educational contexts, they are not especially helpful in gauging or guiding students' learning. As Stake (1991) points out, "The teacher sees education in terms of mastery of specific knowledge and sophistication in the

performance of specific tasks, not in terms of literacy or the many psychological traits commonly defined by our tests." The behaviorist perspective that John Watson introduced in the early 1900s and which remained influential into the 1960s in both theory and instructional practice (e.g., Holland & Skinner, 1961) offers this view of increasing students' capabilities:

> The educational process consists of providing a series of environments that permit the student to learn new behaviors or modify or eliminate existing behaviors and to practice these behaviors to the point that he displays them at some reasonably satisfactory level of competence and regularity under appropriate circumstances. ... The evaluation of the success of instruction and of the student's learning becomes a matter of placing the student in a sample of situations in which the different learned behaviors may appropriately occur and noting the frequency and accuracy with which they do occur.
>
> Krathwohl & Payne, 1971, pp. 17-18

A Toulmin diagram for behaviorist assessments has the same structure as a diagram for trait-based assessments (see Figure 8), with appropriate modifications as to the character of warrants, claims, and data. The psychological/substantive portion of warrants is stimulus/response linkages. Claims concern propensity for the target behavior. One class of data concerns the *features of targeted situations* and the other concerns the *features of performances in those situations*, where the salient features of both are specified in the warrant and defined strictly from the point of view of the assessor. In contrast to Spearman's indifference to the particulars of the situations and behaviors that constituted evidence about a trait, careful attention is focused on specifying situations in behaviorist assessment because behavior in those situations directly defines the characteristic of interest about students. To draw an inference about a student's likely behavior in a domain of such situations, one observes the student's actual behavior in a sample of them. The statistical part of warrant, laid over the substantive part, is a model for success in independent trials—binomial if they are equally difficult, compound binomial if they are not (Lord & Novick, 1968).

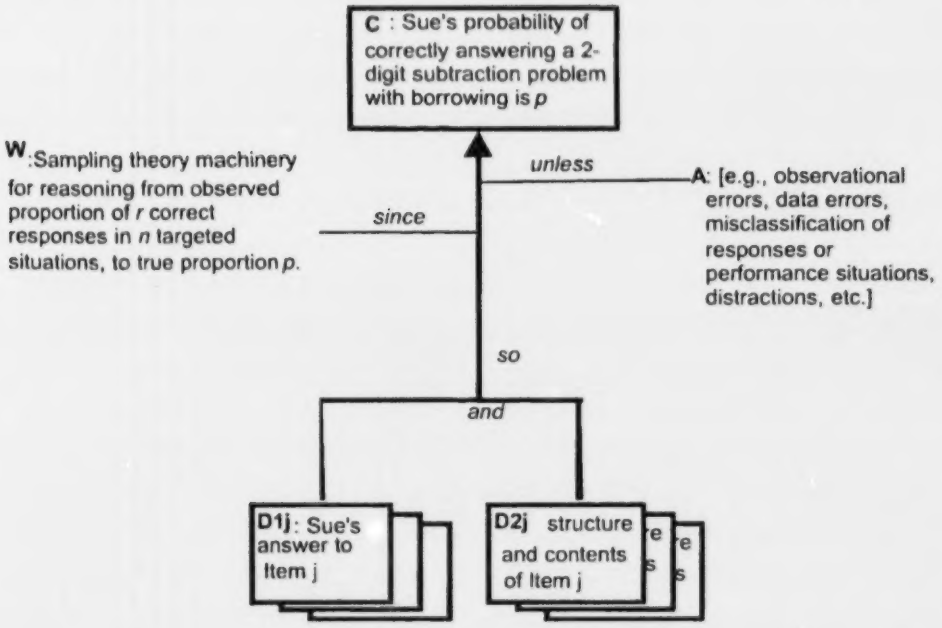

C : Sue's probability of correctly answering a 2-digit subtraction problem with borrowing is p
W:Sampling theory machinery for reasoning from observed proportion of r correct responses in n targeted situations, to true proportion p.
unless
A: [e.g., observational errors, data errors, misclassification of responses or performance situations, distractions, etc.]
since
so
and
D1j: Sue's answer to Item j
D2j structure and contents of Item j

Figure 8. Elaborated Toulmin diagram for multiple observations supporting a behaviorist claim. The warrant encompasses definitions of the class of stimulus situations, response classifications, and sampling theory. The claim addresses only the expected value of performance of the targeted kind in the targeted situations.

Two kinds of tests appeared to support education from this perspective, corresponding to coarser and finer grainsizes. The first, large-scale achievement testing, arose in the 1930s and 1940s to provide measures of relative proficiency in the subjects of school learning, sampled over very broad domains such as science, reading, or mathematics at a given grade level. Covering such a wide span of knowledge and skill in half an hour of testing obviously requires thin sampling, so the results of these tests are not focused enough to guide individual students' instruction. They are meant rather to provide comparable information to determine how well students perform, compared with their grade-level peers in the sampled domains. The second type of test, criterion-referenced tests (CRTs), was introduced in the 1960s as a way of providing instructionally relevant test results to teachers (Glaser, 1963). CRTs address domains defined more narrowly in terms of explicit behavioral objectives. CRTs are designed to estimate students' probabilities of success in a domain, with the goal of determining whether a student has "mastered" it. As with trait-based

assessments, behaviorist assessments do not directly address the processes by which students produce their responses.

Alternative explanations under behaviorist assessment are fairly straightforward because the link between situations, performances, and claims is so direct. They include over- or under-estimating a student's propensity toward the targeted behavior due to incomplete or biased creation of tasks to operationally define the domain and inadequate sampling of tasks. The arguments that are more pertinent are subsequent to the assessment argument itself, regarding the use of these estimated behavioral tendencies. Does a set of propensities toward behavior in domains defined from the assessor's point of view adequately characterize what we want students to know, and does it support our efforts to help them learn it? The answer to these questions, emerging from the so-called "cognitive revolution" in psychology starting in the 1960s, is a resounding no.

Assessment Arguments Under the Information-Processing Perspective

Like behavioral psychologists, cognitive psychologists who are interested in learning attend to the features of situations in which knowledge is acquired, and the contexts in which people use it. Analysis and decomposition of features of situations may again be employed. The information-processing view goes further, though, by taking internal representations of the situation and the behavior as targets of study. In ways both conscious and subconscious, the task a student solves is not the problem as the investigator poses it, but the problem as the student perceives it. Studies contrasting experts and novices in domains as diverse as chess (de Groot, 1965), radiology (Lesgold, Feltovic, Glaser, & Wang, 1981), writing (Scarmadelia & Berieter, 1991), and volleyball (Allard & Starkes, 1980) reveal variations on a common theme. In each case, in a way that reflects the domain, experts

> ... (a) provide coherent explanations based on underlying principles rather than descriptions of superficial features or single statements of fact, (b) generate a plan for solution that is guided by an adequate representation of the problem situation and possible procedures and outcomes, (c) implement solution strategies that reflect relevant goals and subgoals, and (d) monitor their actions and flexibly adjust their approach based on performance feedback.
>
> Baxter, Elder, & Glaser, 1996, p. 133.

The claims of interest in assessment designed from an information-processing perspective, then, are not merely patterns of students' behavior in situations with features that are salient from the assessor's point of view. Rather, claims concern knowledge structures, mappings of situations into knowledge structures, and performance in situations as it is mediated by those knowledge structures. Patterns of actions in suitably defined task situations still provide evidence about behavioral propensities, but this is now only an intermediate stage in an assessment argument about a student's cognition. The central inferential question is now, as Thompson (1982) put it, "What can this person be thinking so that his actions make sense from his perspective?" The importance of assessments cast from an information-processing perspective is the more direct connection between claims and instruction. That is, inferences are organized directly in terms of the underlying concepts, relationships, and strategies for tackling problems in the domain, rather than indirectly in terms of features of problems as an expert sees them.

Brown and Burton (1978), for example, analyzed children's subtraction in terms of the set of so-called production rules—some correct, some perhaps buggy—that students could bring to bear on problems. Claims here were in terms of production rules hypothesized to govern a student's solutions. The warrant was in terms of the responses—some correct, sometimes for the wrong reasons, some incorrect, with answers that reflected buggy rules—that would likely be produced by a student with a given set of production rules. Figure 9 provides an example. Domains of tasks can still be grouped by features that are similar from the assessor's point of view, but the target of inference is the student's thinking that make them similar from his or her point of view. A Toulmin diagram that corresponds to Brown and Burton's assessment argument is shown as Figure 10. The lower part of the diagram is much like that of a number of assessments cast under a behaviorist perspective: Data consist of aspects of students' actions and features of situations, arising from some propensities to such behavior. The higher level, however, is a (possibly multifaceted) claim about the knowledge representations through which the student has perceived the situations—possibly quite different from the assessor's—and the procedures and strategies the student brings to bear on problems as he or she perceives them.

821 - 285 664	885 - 221 664
63 - 15 52	17 - 9 12

Figure 9. Responses consistent with the "subtract smaller from larger" bug. When the 'subtract smaller from larger' bug is present in a student's configuration of production rules, problems requiring borrowing will show the characteristic pattern of incorrect responses that results from simply subtracting whichever number in a column is smaller from whichever is larger. When borrowing is not required, this bug does not affect responses; they will be correct or incorrect in whatever ways are consistent with the student's other rules.

Sparked by John B. Carroll's (1976) pioneering studies, an active area of research on assessment is exploiting what can be learned from information-processing analyses of tasks in several ways. Warrants are cast explicitly in information processing terms. Task features, one portion of the assessment data, are designed around features suggested by the theory of the domain (e.g., Embretson, 1998). Student performances, another portion of the data, are evaluated in terms of behaviors suggested by the theory of the domain. For example, Mosenthal and Kirsch's (1991) cognitive model for processing documents indicates that the difficulty of a task will be driven by (a) features of the document in question, such as the number of organizing categories, (b) features of the directive, such the number of features that must be matched, and (c) the correspondence between the two, as determined by the degree to which the document has been designed to facilitate the inference that must be drawn.

And psychometric models have been introduced to handle claims cast in information-processing terms, explicitly modeling performance in terms of theory-based predictions of performance (see Junker, 1999, for a recent review). Referring back to the analytical reasoning items introduced in the section on trait-based assessment, cognitive analyses of solutions of items of this type have led to grammars for describing features of such items; for manipulating features to make them harder or easier by increasing or decreasing their loads on working

memory, representational form, or contextual knowledge; and for modeling their operating characteristics in psychometric models in terms of their cognitively salient features. Indeed, Newstead, Bradon, Handley, Evans, & Dennis (2002) propose such a model for analytic reasoning items like Pet Shop, where the salient features include the number and type of constraints and how many configurations would be compatible with the stated requirements. (The interested reader is referred to Embretson [1998] on the coordination of among cognitive theory, task design, and psychometric modeling in psychological assessment.)

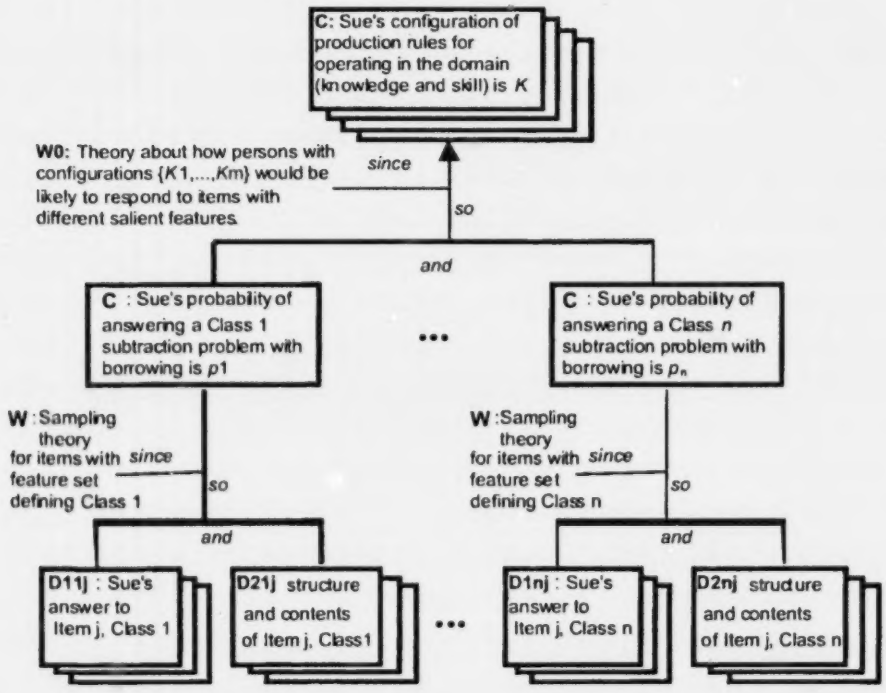

C: Sue's configuration of production rules for operating in the domain (knowledge and skill) is K
W0: Theory about how persons with configurations {K1,...,Km} would be likely to respond to items with different salient features
since
so
and
C : Sue's probability of answering a Class 1 subtraction problem with borrowing is p1
C : Sue's probability of answering a Class n subtraction problem with borrowing is pn
W : Sampling theory for items with feature set defining Class 1
since
so
W : Sampling theory for items with feature set defining Class n
since
so
and
and
D11j : Sue's answer to Item j, Class 1
D21j structure and contents of Item j, Class1
D1nj : Sue's answer to Item j, Class n
D2nj structure and contents of Item j, Class n

Figure 10. Elaborated Toulmin diagram for inference about cognitive model in the domain of whole number subtraction. Responses and performance situations can be identical to collections of those used in a series of behavioral assessments about performance in categories of items. However, ultimate claim is characterization of the student in terms of recognizing structures of problems and having skills and strategies to apply to solve them. Behavior across patterns of problems of different classes is evidence for the underlying set of rules that characterize the student.

BEST COPY AVAILABLE

Another aspect of competence that has emerged from studies of expertise is the iterative character of complex problem solving. Both scientists engaged in inquiry and mechanics fixing hydraulics systems generate hypotheses and provisional models, take actions to test them, and revise their understanding to proceed to the next step (White & Frederiksen, 1998). This is a modeling challenge for assessment because the information from different time points is serially independent. At each time point, the performance situation changes as a result of the examinee's previous actions and their effects on the system, as suggested in Figure 11. Furthermore, the evaluation of actions at each time point

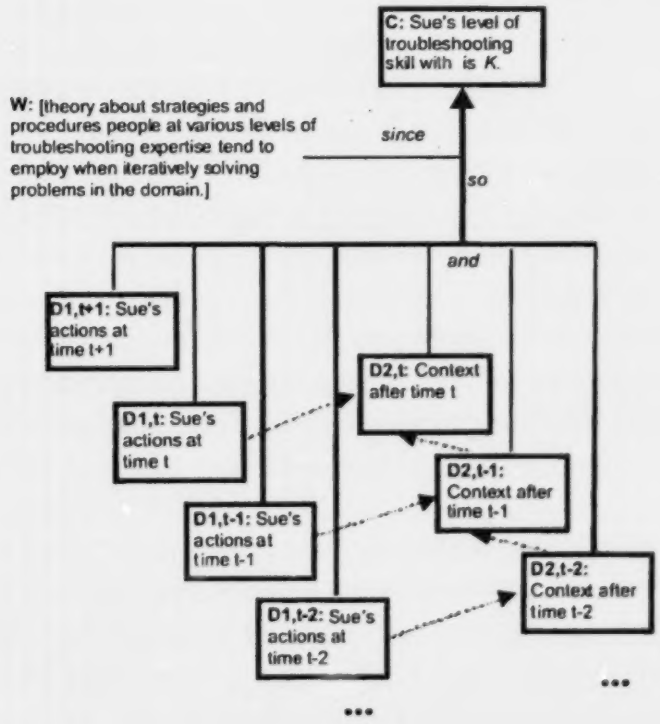

C: Sue's level of troubleshooting skill with is K.
W: [theory about strategies and procedures people at various levels of troubleshooting expertise tend to employ when iteratively solving problems in the domain.]
since
so
and
D1,t+1: Sue's actions at time t+1
D2,t: Context after time t
D1,t: Sue's actions at time t
D2,t-1: Context after time t-1
D1,t-1: Sue's actions at time t-1
D2,t-2: Context after time t-2
D1,t-2: Sue's actions at time t-2

Figure 11. Elaborated Toulmin diagram for assessing hydraulics system troubleshooting. Direct evidence for a claim about troubleshooting competence is obtained with a sequence of interactions between the examinee and the system, as the examinee proceeds through situation-hypothesis-action cycles. At each time point, the performance situation changes as a result of the examinee's previous actions and their effects on the system. Therefore, the data from actions at different time points are serially independent.

must take into account not only the immediate action, but the situation as it has evolved thus far and the informational relationship of this action to previous actions. The same test of the same valve in a hydraulics problem can reflect the expert-level space-splitting if performed early, but it can be redundant if performed after a different test has already eliminated that part of the system as the source of the fault.

Assessment Arguments Under the Sociocultural Perspective

Much learning is motivated and evaluated, then, by the knowledge, goals, constraints, and physical presence of other people. Social organizations such as families, classrooms, professions, and so on, influence the processes of acquiring, storing, representing, understanding, and creating knowledge. Moreover, many of these influences are channeled by particular ways of communicating; knowledge representations, genres, conventions, and so on. From the sociocultural point of view, knowledge is developed through the practical activities of groups of people as they interact in various contexts with each other and with resources such as books and tools. "Learning by a group or individual involves becoming attuned to the constraints and affordances of material and social systems with which they interact" (Greeno, et al., 1997, p. 17).

The sociocultural perspective proposes a view on the nature of knowledge and learning, and consequently on the nature of warrants, claims, and data for assessment framed under its aegis. In particular, "The situated view of assessment emphasizes questions about the quality of student participation in activities of inquiry and sense making, and considers assessment practices as integral components of the general systems of activity in which they occur" (Greeno, et al., 1997, p. 37). Compared to the information-processing perspective, there is a greater emphasis on patterns of interactions of students with people and social artifacts and less emphasis on knowledge structures "inside the students' heads."

The intimate connection between features of situations for acquiring and using knowledge on the one hand, and features of situations necessary for obtaining evidence about that knowledge, adds a layer of complexity to assessment under the sociocultural perspective. Contextualizing assessment decreases the assessor's control over the features of the observational situation. It increases the burden on arranging for and identifying the salient features of both

performances and performance situations. It introduces alternative explanations for good and poor performance, in connection with characteristics of the situations, people, and materials with whom the assessed student interacts. The challenges are not insurmountable, however. Work by Wiggins (1998) and White and Frederiksen (2000) on designing assessment to produce the kinds of learning that are valued under the sociocultural perspective is at once grounded in sound evidentiary reasoning and practical for classroom use. The two following examples illustrate key issues that arise in assessment from a sociocultural perspective.

Example: Advanced Placement Studio Art Portfolio Assessment. The purpose of the College Entrance Examination Board's Advanced Placement (AP) Studio Art portfolio assessment is to determine whether high school students exhibit knowledge and skills commensurate with first-year post-secondary art courses (Mitchell, 1992). Students develop works for their portfolios in their local classes during the course of the year, through which they demonstrate the knowledge and skills described in the AP Studio Art materials. The portfolios are rated centrally by artist/educators at the end of the year, using standards set in general terms and monitored by the AP Art advisory committee. These standards are rendered in language sufficiently general to apply to a wide range of subjects, styles, and media. Their meaning is constructed over time and across sites through shared examples, not unlike the way the meaning of a law evolves as it is applied to particular cases. Coming to learn this language, this artist's way of seeing the world, in evaluating their own work and that of others is in fact a key learning goal of the program. Assessment here concerns "questions of what is of value, rather than simple correctness ... an episode in which students and teachers might learn, through reflection and debate, about the standards of good work and the rules of evidence" (Wolf, Bixby, Glenn, & Gardner, 1991, p. 51).

Section B of the portfolio, the student's "concentration," is of particular interest in regard to constructing a warrant and reasoning through it from student work to ratings. A concentration consists of up to 20 slides, a film, or a videotape illustrating extended work on a student-selected theme, and a narrative describing the student's goals, intentions, influences, and other factors that help explain the work. In the narrative, the student makes a case for how the common rubric for evaluating concentrations should be applied to her particular work. Not only the data but also the application of the warrant itself

are negotiated between the student and the assessor. The "other factors" in the narrative further serve to supply information to deal with alternative explanations for good or poor performance, such as pointing to a trend in the work over time that mitigates the lower quality of earlier works. Of course this data itself introduces issues of sensitivity and objectivity on the part of the student supplying it. Figure 12 suggests how the student narrative functions both as data which is evaluated and backing for a warrant as it is tailored to interpreting the artworks.

A psychometric model is used in AP Studio Art to combine the scores that several raters assign to several sections of a student's portfolio. However, the student variable is not interpreted as a trait, as it would have been under the differential/trait perspective under which the model originated. Rather, the score is a synthesis of informed and socially moderated judgments about a

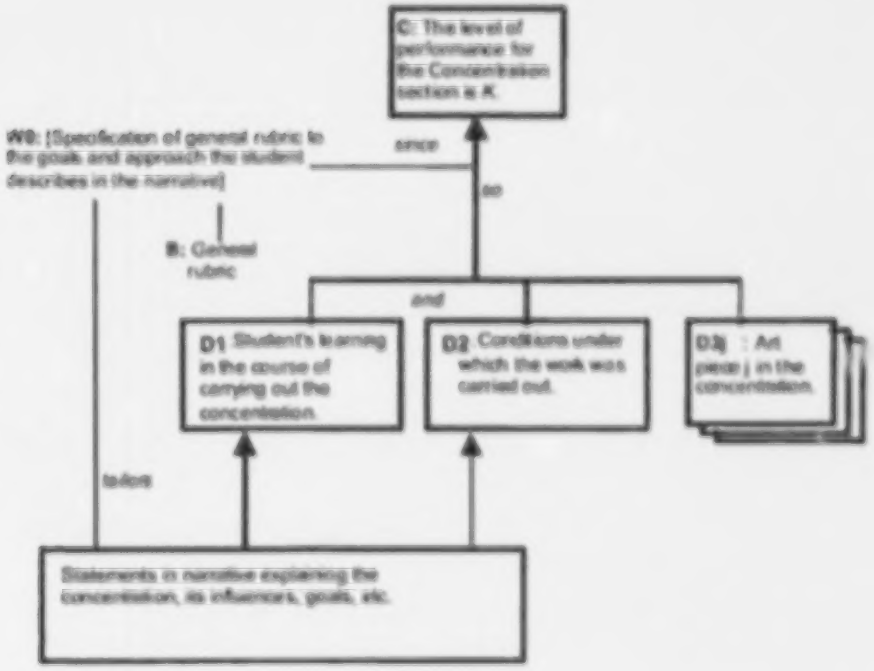

C: The level of performance for the Concentration section is K.
W0: [Specification of general rubric to the goals and approach the student describes in the narrative]
since
so
B: General rubric
and
D1 Student's learning in the course of carrying out the concentration.
D2 Conditions under which the work was carried out.
D3j : Art piece j in the concentration.
tailors
Statements in narrative explaining the concentration, its influences, goals, etc.

Figure 12. Elaborated Toulmin diagram for Advanced Placement Studio Art portfolio assessment. Statements in the narrative contribute to knowledge about the thinking behind the student's art works, the conditions under which the work was produced, and the application of the generally stated rubric to the work.

BEST COPY AVAILABLE

student's specific accomplishments. The psychometric model is not a tool for measuring qualities of students, but a way to manage episodes of evaluation across many portfolios, many raters, and many portfolio sections. Importantly, one of its key roles is to identify portfolios that provoke discrepant scores from different raters. These often turn out to be works that are unique in perspectives or techniques. The extended discussions that are accorded to these portfolios yield a final judgment, as they must, but extend the meaning of the evaluation procedures for raters and future students alike.

Example: Conversational competence. The most familiar form of large-scale language testing addresses the knowledge of language per se, exercising points of vocabulary, syntax, and comprehension with discrete and largely decontextualized test items. Assessments so constructed fit comfortably into trait, and sometimes behavioral, perspectives on learning, and support assessment arguments framed in their terms. This kind of knowledge is not enough to use a language to achieve ends in social situations. In addition to grammatical competence, we must be concerned with the social context of language use, pragmatic considerations in using language to achieve goals, and familiarity with forms, customs, and standards of communication above the level of sentences. Conversational competence (Widdowson, 1978), for example, means being able to participate in an interaction with another person in using all of these kinds of knowledge to construct a joint understanding of a situation of mutual interest to achieve some purpose.

Obtaining direct evidence about conversational competence, then, requires observing a student engaging in the interaction that characterizes conversation. Four factors immediately impact the assessment argument. First, the events in conversations are not conditionally independent given the participants' conversational competence, but are serially dependent. In addition to whatever global context a conversation occurs in, the local context for every utterance depends on what the speaker and the interlocutors have already said. Second, more than one person is acting, and each person's actions influence and set the performance context for the other. A Toulmin diagram that suggests these first two factors is shown as Figure 13. Third, the rules by which students' performances are evaluated must take these dependencies and interactions into account. The dependencies are similar to the ones in dynamic problem-solving discussed above in connection with the information-processing perspective; we

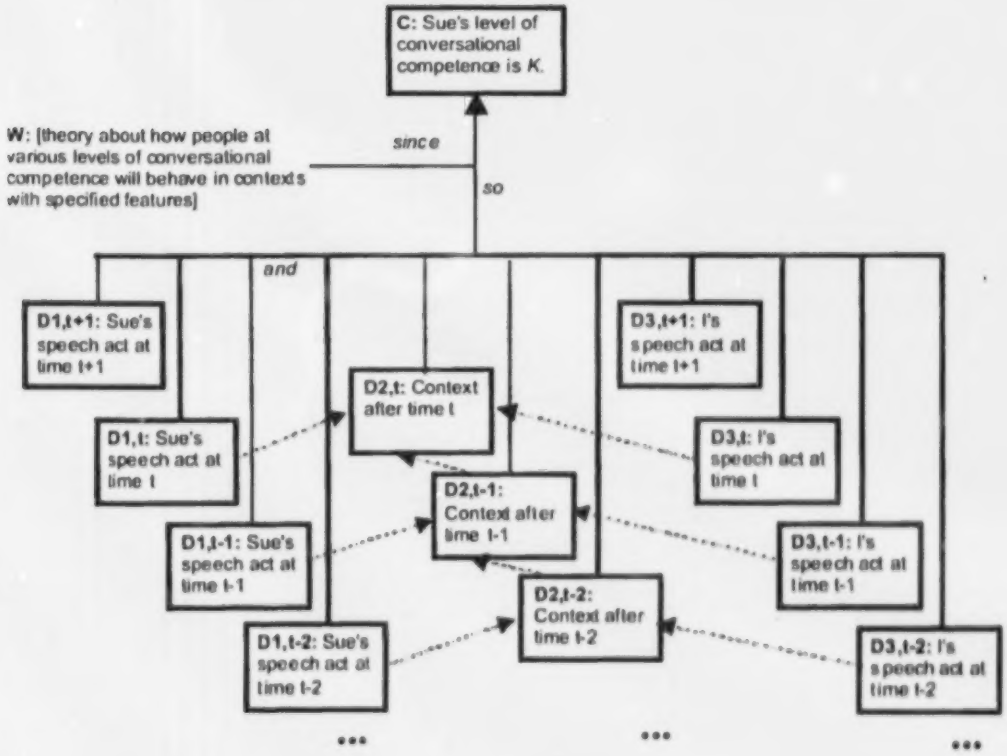

C: Sue's level of conversational competence is K.
W: [theory about how people at various levels of conversational competence will behave in contexts with specified features]
since
so
and
D1,t+1: Sue's speech act at time t+1
D1,t: Sue's speech act at time t
D1,t-1: Sue's speech act at time t-1
D1,t-2: Sue's speech act at time t-2
D2,t: Context after time t
D2,t-1: Context after time t-1
D2,t-2: Context after time t-2
D3,t+1: I's speech act at time t+1
D3,t: I's speech act at time t
D3,t-1: I's speech act at time t-1
D3,t-2: I's speech act at time t-2

Figure 13. Elaborated Toulmin diagram for assessing conversational competence. Direct evidence for a claim about conversational competence is obtained through interactions between two or more people—two are addressed here. At each time point, the utterances of a person become part of a common performance situation, the context within which the next action must be evaluated. This figure concerns an oral interview in which only a claim about the student's competence is desired.

have now added the complexities that arise when two people use interpersonal and cultural knowledge in addition to domain knowledge, to jointly achieve some goal.

The fourth factor is that as local contexts are constructed, the conditions necessary for observing a certain facet of conversational competence—switching from formal to informal register, for example—may not emerge. In contrast with the trait-, behavioral-, and information-processing perspectives, an assessment designer working under the sociocultural perspective has less control over the

contextual data if the salient features of a performance situation can only emerge from interactions among individuals. As an assessment task, a minimally constrained conversation between two students can have the advantage of meaningfulness to the participants. This mitigates alternative explanations of poor performance that stem from lack of background knowledge or motivation. But the need to switch registers may not arise, and the observation provides no evidence about a targeted facet of ability. On the other hand, a trained interviewer can guide a conversation in a way that provokes register-switching, facets of language, or social conventions. The performance situation features which, under the conversational competence warrant, are needed to obtain evidence about targeted facts of ability can be better assured—though now at the cost of introducing alternative explanations for poor performance based on the artificiality of the conversation. This is an example of the fundamental tradeoffs in assessment design. Every performance situation taps myriad aspects of knowledge and ability, and choosing a configuration that counters one alternative explanation inevitably opens the door for another.

The effective meaning of student variables in such an assessment arises from the features of the situations in which students perform and the features of their actions that are evaluated. The degree to which inferences about students based on observations in these particular settings (including the participants and contexts) will apply to other settings is a matter of similarities to other potential settings, and students' tendencies to interact in those settings in similar or different ways. This is fundamentally the same question of the generalizability of test scores that arises under trait-based interpretations. In both cases it is ultimately an empirical question. How far will the meaning of scores generalize? Before the data are in, the trait psychologist's prior is "pretty far"; the sociocultural psychologist thinks, "I wouldn't bet on it."

Discussion

The forms and the uses of educational assessment have changed considerably over the past century in response to changing views of the nature of knowledge—how it is acquired, how schooling should be organized to promote it, and how assessments should be designed to guide instruction. Continued changes, perhaps ultimately even more radical, are taking place today as a result of new technologies for gathering and analyzing performance data. A closer look

at the structure of the arguments beneath assessments that can appear very different on the surface reveals a deeper kind of stability. The stability is found in terms of the core argument of any assessment: We want to draw inferences about what students know and can do as seen from some perspective; that perspective tells us what kinds of things we need to see them do, in what kinds of situations, to ground those inferences. We see elaborations, extensions, and specializations of enduring principles of evidentiary reasoning. We find continued value in knowledge representations such as Toulmin diagrams, Wigmore charts, and Bayesian inference networks to understand yesterday's assessments, manage today's, and design the assessments of tomorrow.

References

Allard, F. & Starkes, J. L. (1980). Perception in sport: Volleyball. *Journal of Sport Psychology, 2,* 22-23.

Almond, R. G., Steinberg, L. S., & Mislevy, R. J. (2002). Enhancing the design and delivery of assessment systems: A four-process architecture. *Journal of Technology, Learning, and Assessment, 1*(5). http://www.bc.edu/research/intasc/jtla/journal/v1n5.shtml

Anderson, T. J., & Twining, W. L. (1991). *Analysis of evidence.* Boston: Little, Brown, & Co.

Baxter, G. P., Elder, A. D., & Glaser, R., (1996). Knowledge-based cognition and performance assessment in the science classroom. *Educational Psychologist, 31* (2), 133-140.

Brown, J. S., & Burton, R. R. (1978). Diagnostic models for procedural errors in basic mathematical skills. *Cognitive Science, 2,* 155-192.

Carroll, J. B. (1976). Psychometric tests as cognitive tasks: A new "structure of intellect." In L.B. Resnick (Ed.), *The nature of intelligence* (pp. 27-56). Hillsdale, NJ: Erlbaum.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity and psychological tests. *Psychological Bulletin, 52,* 281-302.

de Groot, A. D. (1965). *Thought and choice in chess.* The Hague: Mouton.

Edwards, W. (1998). Hailfinder: Tools for and experiences with Bayesian normative modeling. *American Psychologist, 53,* 416-428.

Embretson, S. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin, 93,* 179-197.

Embretson, S. E. (1998). A cognitive design system approach to generating valid tests: Application to abstract reasoning. *Psychological Methods, 3,* 380-396.

French, J. W. (1965). The relationship of problem-solving styles to the factor composition of tests. *Educational and Psychological Measurement, 25,* 9-28.

Glaser, R. (1963). Instructional technology and the measurement of learning outcomes: Some questions. *American Psychologist, 118,* 519-521.

Glaser, R. (1981). The future of testing: A research agenda for cognitive psychology and psychometrics. *American Psychologist, 36,* 923-936.

Gould, S. J. (1981). *The mismeasure of man*. Harmondsworth: Penguins Books.

Greeno, J. G., Collins, A. M., & Resnick, L. B. (1997). Cognition and learning. In D. Berliner & R. Calfee (Eds.), *Handbook of educational psychology* (pp. 15-47). New York: Simon & Schuster Macmillan.

Gulliksen, H. (1961). Measurement of learning and mental abilities. *Psychometrika, 26,* 93-107.

Holland, J. G., & Skinner, B. F. (1961). *The analysis of behavior: A program for self-instruction.* New York: McGraw-Hill.

Jensen, F. V. (1996). *An introduction to Bayesian networks.* New York: Springer-Verlag.

Junker, B. J. (1999). Some statistical models and computational methods that may be useful for cognitively relevant assessment. Commissioned paper prepared for the Committee on the Foundations of Assessment, National Research Council. Retrieved September 5, 2003 from http://www.stat.cmu.edu/~brian/nrc/cfa/

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin, 112,* 527-535.

Krathwohl, D. R., & Payne, D. A. (1971). Defining and assessing educational objectives. In R. L. Thorndike (Ed.), *Educational measurement* (2nd Ed.) (pp. 17-45). Washington, DC: American Council on Education.

Lewis, C. (1986). Test theory and *Psychometrika*: The past twenty-five years. *Psychometrika, 51,* 11-22.

Lesgold, A. M., Feltovich, P. J., Glaser, R., & Wang, Y. (1981). The acquisition of perceptual diagnostic skill in radiology. *Technical Report No. PDS-1.* Pittsburgh: Learning Research and Development Center, University of Pittsburgh.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York: American Council on Education/Macmillan.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Education Researcher,* 32(2), 13-23.

Mislevy, R. J. (1994). Evidence and inference in educational assessment. *Psychometrika, 59,* 439-483.

Mislevy, R. J., & Gitomer, D. H. (1996). The role of probability-based inference in an intelligent tutoring system. *User-Modeling and User-Adapted Interaction, 5,* 253-282.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives, 1,* 3-67.

Mislevy, R. J., Wilson, M. R., Ercikan, K., & Chudowsky, N. (2003). Psychometric principles in student assessment. In T. Kellaghan & D. Stufflebeam (Eds.), *International Handbook of Educational Evaluation* (pp. 489-531). Dordrecht, the Netherlands: Kluwer Academic Press.

Mitchell, R. (1992). *Testing for learning: How new approaches to evaluation can improve American schools.* New York: The Free Press.

Mosenthal, P. B., & Kirsch, I. S. (1991). Toward an explanatory model of document process. *Discourse Processes, 14,*147-180.

National Research Council (2001). *Knowing what students know: The science and design of educational assessment.* Committee on the Foundations of Assessment, J. Pellegrino, R. Glaser, & N. Chudowsky (Eds.). Washington DC: National Academy Press.

Newell, A., & Simon, H.A. (1972). *Human problem solving.* Englewood Cliffs, NJ: Prentice-Hall.

Newstead, S., Bradon, P., Handley, S., Evans, J., & Dennis, I. (2002). Using the psychology of reasoning to predict the difficulty of analytical reasoning problems. In S.H. Irvine & P.C. Kyllonen (Eds.), *Item generation for test development* (pp. 35-52). Mahwah, NJ: Erlbaum.

Scardamalia, M., & Bereiter. (1991). Literate expertise. In K. A. Ericcson & J. Smith (Eds.), *Toward a general theory of expertise* (pp. 172-194). Cambridge, England: Cambridge University Press.

Schum, D. A. (1987). *Evidence and inference for the intelligence analyst.* Lanham, MD: University Press of America.

Schum, D. A. (1994). *The evidential foundations of probabilistic reasoning.* New York: Wiley.

Spearman, C. (1927). *The abilities of man: Their nature and measurement.* New York: Macmillan,

Stake, R. E. (1991). The teacher, standardized testing, and prospects of revolution. *Phi Delta Kappan, 73,* 243-247.

Thompson, P. W. (1982). Were lions to speak, we wouldn't understand. *Journal of Mathematical Behavior, 3,* 147-165.

Tillers, P., & Schum, D. A. (1991). A theory of preliminary fact investigation. *UC Davis Law Review, 24,* 907-966.

Toulmin, S. E. (1958). *The uses of argument.* Cambridge: Cambridge University Press.

Wainer, H., Dorans, N. J., Flaugher, R., Green, B. F., Mislevy, R. J., Steinberg, L., et al. (2000). *Computerized adaptive testing: A primer* (second edition). Hillsdale, NJ: Lawrence Erlbaum Associates.

White, B., & Frederiksen, J. (1998) Inquiry, Modeling, and Metacognition: Making science accessible to all students. *Cognition and Instruction, 16,* 3-118.

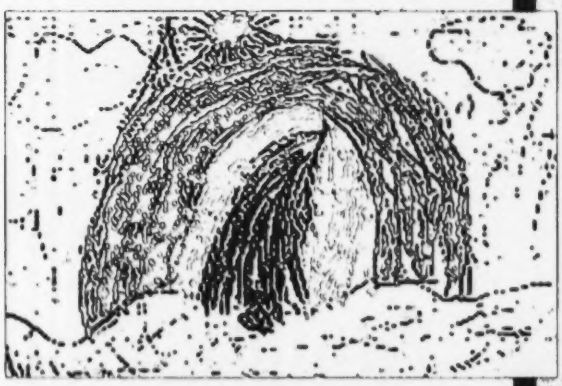
White, B. Y., & Frederiksen, J. R. (2000). Metacognitive facilitation: An approach to making scientific inquiry accessible to all. In J. Minstrell & E. van Zee (Eds.), *Teaching in the inquiry-based science classroom.* Washington, DC: American Association for the Advancement of Science.

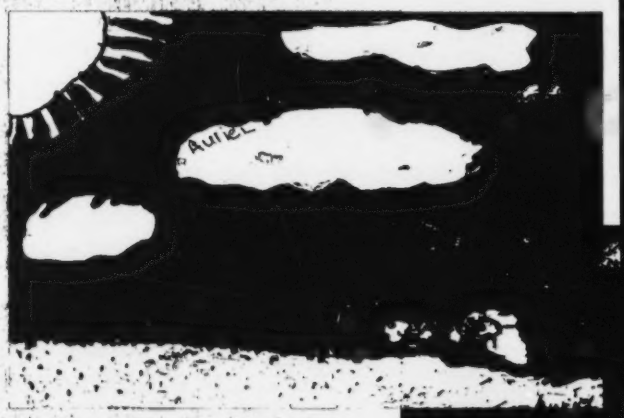
Widdowson, H. G. 1978. *Teaching language and communication.* Oxford: Oxford University Press.

Wiggins, G. P. (1998). *Educative assessment: Designing assessments to inform and improve student performance.* San Francisco: Jossey-Bass.

Wigmore, J. H. (1937). *The science of judicial proof* (3rd Ed.). Boston: Little, Brown, & Co.

Wolf, D., Bixby, J., Glenn, J., & Gardner, H. (1991). To use their minds well: Investigating new forms of student assessment. In G. Grant (Ed.), *Review of Educational Research, Vol. 17* (pp. 31-74). Washington, DC: American Educational Research Association.





BEST COPY AVAILABLE

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

NOTICE

Reproduction Basis

[X] This document is covered by a signed "Reproduction Release (Blanket)" form (on file within the ERIC system), encompassing all or classes of documents from its source organization and, therefore, does not require a "Specific Document" Release form.

[ ] This document is Federally-funded, or carries its own permission to reproduce, or is otherwise in the public domain and, therefore, may be reproduced by ERIC without a signed Reproduction Release form (either "Specific Document" or "Blanket").